Chapter 3

Use of Different Sources of Information in Maintaining Standards: Examples from the Netherlands

Anton Béguin

Abstract In the different tests and examinations that are used at a national level in the Netherlands, a variety of equating and linking procedures are applied to maintain assessment standards. This chapter presents an overview of potential sources of information that can be used in the standard setting of tests and examinations. Examples from test practices in the Netherlands are provided that apply some of these sources of information. This chapter discusses how the different sources of information are applied and aggregated to set the levels. It also discusses under which circumstances performance information of the population would be sufficient to set the levels and when additional information is necessary.

Keywords: linking, random equivalent groups equating, nonequivalent groups equating

Introduction

In the different tests and examinations that are used at a national level in the Netherlands, a variety of equating and linking procedures are applied to maintain assessment standards. Three different types of approaches can be distinguished. First, equated scores are determined to compare a new form of a test to an existing form, based on an anchor that provides information on how the two tests relate in difficulty level and potentially in other statistical characteristics. A special version of this equating procedure is applied in the construction and application of item banks, in which the setting of the cut-score of a test form is based on the underlying Item Response Theory (IRT) scale. Second, in certain instances—for example, central examinations at the end of secondary education—heuristic procedures are developed to incorporate different sources of information, such as pretest and anchor test data, qualitative judgments about the difficulty level of a test, and the development over time of the proficiency level of the population. For each source of the data, the optimal cut-scores on the test are determined. Because the validity of assumptions and the accuracy of the data are crucial factors, confidence intervals around the cut-scores are determined, and a heuristic is applied to aggregate the results from the different data sources.

Third, in the standard setting of a test at the end of primary education, significant weight is assigned to the assumption of random equivalent groups, whereas the other sources of information (pretest data, results on similar tests, and anchor information) are mainly used as a check on the validity of the equating. In the current chapter, an overview of potential sources of information that can be used in the standard setting of examinations is presented. The overview includes information on the following:

1. Linking data that can be used in equating and IRT linking procedures, with various data collection designs and different statistical procedures available
2. Different types of qualitative judgments: estimates of difficulty level/estimates of performance level
3. Assumptions made in relation to equivalent populations
4. The prior performance of the same students
5. The historical difficulty level of the test forms

Examples from test practices in the Netherlands are provided that apply some of these sources of information, which are then aggregated and applied in the standard-setting procedure to set the levels. This chapter discusses the advantages and disadvantages of some of the sources of information, especially regarding under which circumstances random equivalent groups equating—using only performance information of the population—would improve the quality or efficiency of the level-setting procedure and when additional information is necessary.

Sources of Information for Standard Setting

Linking Data

To be able to compare different forms of a test, one needs either linking data or an assumption of random equivalent groups. A number of different designs and data collection procedures have been distinguished (Angoff, 1971; Béguin, 2000; Holland & Dorans, 2006; Kolen & Brennan, 2004; Lord, 1980; Petersen, Kolen, & Hoover, 1989; Wright & Stone, 1979). In these data collection procedures, a distinction can be drawn between designs that assume that the test forms are administered to a single group or to random equivalent groups and nonequivalent group designs for which the assumption of random equivalent groups may not hold. In the context of examinations, the data collected during actual exams can theoretically be treated as data from a random equivalent groups design.

Each form of the examination is administered to separate groups of respondents, but it is assumed that these groups are randomly equivalent. More relevant in the current context are

the nonequivalent groups designs. Examples of such designs are anchor test designs, designs using embedded items, and pretest designs.

A variety of equating procedures are available to compare test forms. These procedures use the collected data to estimate the performance characteristics of a single group on a number of different test forms (e.g., to estimate which scores are equivalent between forms and how cut-scores can be translated from one form to the other).

The equating procedures either use only observable variables or assume latent variables, such as a true score or a latent proficiency variable. Procedures using only observable variables are, for example, the Tucker method (Gulliksen, 1950), Braun-Holland method (Braun & Holland, 1982), and chained equipercentile method (Angoff, 1971). Latent variable procedures include Levine's (1955) linear true score equating procedure and various procedures based on IRT (e.g., Kolen & Brennan, 2004; Lord, 1980).

Item Banking

Item-banking procedures can be considered a special case of equating procedures. These procedures often use a complex design to link new items to an existing item bank. They rely heavily on statistical models, the assumption that the characteristics of items can be estimated, and that these characteristics remain stable during at least a period of time. Typically, item banks are maintained by embedding new items within live test versions or by the administration of a separate pretest. If an IRT model is used, often parameters for difficulty, discrimination, and guessing are estimated. To ensure that the new items are on the same scale as the items in the bank, the new items are calibrated, together with the items for which item characteristics are available in the bank. To evaluate whether the above procedure is valid, it is crucial that the underlying assumptions are checked. For example, the stability of the items' characteristics needs to be evaluated, comparing between the previous administrations on which the item characteristics are based and the performance in the current administration. The stability can be violated in cases where items are administered under time constraints, order effects occur, or if items become known due to previous administrations.

Because of the potential adverse effect of these issues on the validity of the equating, it is crucial to monitor the performance of the individual items and the validity of the link between the new test version and the item bank.

Clearly, the variables of interest in level setting, such as equivalent cut-scores, are directly affected by the quality and the stability of the equating procedure. The quality of the equating of test forms largely depends on potential threads to validity in the data collection. For

example, the results of a pretest could potentially be biased if order effects and administration effects are not dealt with appropriately. The stability of equating depends on the quality and the size of the sample, characteristics of the data collection design, and the equating procedure that is used (e.g., Hanson & Béguin, 2002; Kim & Cohen, 1998).

Qualitative Judgments

To set cut-scores on a test form, standard setting procedures based on qualitative judgments about the difficulty level of the test form can be applied. Various procedures are available (e.g., Cizek, 1996, 2001; Hambleton & Pitoniak, 2006), ranging from purely content-based procedures (Angoff procedure, bookmark procedure), which focus on the content of the test, to candidate-centered procedures (borderline, contrasting groups), which aim to estimate a cut-score based on differences between groups of candidates. For example, in a contrasting-groups procedure, raters are asked to distinguish between groups of candidates who perform below the level necessary to pass the test and groups of candidates who perform above this level. In this judgment, the raters do not use the test score. Then the test score distributions of these groups are contrasted to select the cut-score that best distinguishes between the two groups.

The quality of a level-setting procedure largely depends on the quality of the judges, the number of judges involved, the characteristics of the procedure, and the quality of the instruction. Often, relatively unstable or biased results are obtained in cases where the instruction or the number of judges is insufficient.

Random Equivalent Groups

In contrast to many other sources of equating information, the performance level of the population is often a very stable measure. Comparing the performance level of the population between one year and the next will only result in large differences if the composition of the population or the curriculum has changed. Differences in year-to-year performance could also occur if there is an increasing or decreasing trend in performance. However, in a number of cases, it is not unreasonable to make an assumption of random equivalent groups from one year to the next. Based on this assumption, it is possible to apply level-setting procedures.

An extended version of the assumption of random equivalent groups takes background variables into account. If the year-to-year populations differ in composition based on a number of background variables, this difference can be corrected using weighing. In such cases, groups of students with the same background variable are assumed to be a random

sample from the same population. Using weighing based on background variables, the assumption of random equivalent groups will hold again in the total population.

Prior Performance of the Same Group

Procedures used to estimate the performance level based on prior attainment on a test a few years earlier can be viewed as a special case of taking background information into account. Two pieces of information can be derived from the prior attainment data: On one hand, the data show whether the population deviates from the average. A correction for this would be similar to the extended assumption of random equivalent groups described above. On the other hand, the prior attainment data could provide information on the performance levels that were reached earlier. Using the information on how the prior performance relates to the standards on the new test form, the cut-scores on this new form can be estimated.

Historical Difficulty Level of the Test Forms

The variation in the difficulty level of the test forms constructed according to the same test blueprint can be used to estimate the difficulty of the current test form. Assuming that the current test form will not be significantly different from the previous forms (e.g., over the past 10 years) will result in a confidence interval. Using historical information, it is assumed that the difficulty of this year's form will fall within this confidence interval.

Linking Procedures Used in Some of the Principal Tests in the Netherlands

Entrance Test to Teacher Training

During the first year of the teacher training program, students have to pass tests in mathematics and in the Dutch language. Students will have a maximum of three opportunities to pass these tests. If they fail these attempts, they are not allowed to continue their education.

The mathematics test is an adaptive test based on an underlying item bank calibrated using the one-parameter logistic model (OPLM) (Verhelst, Glas, & Verstralen, 1994). The item parameters are based on samples with at least 600 respondents for each item in the bank. In addition to the data on the respondents from the teacher training program, these samples may also include information collected from other fields of education.

The bank may contain, for example, items that originated in primary education. In such cases, the original item parameters are based on the performance of students in primary education. On a yearly basis, the parameters are updated based on the performance during the actual

administration of the test. New items are pretested on a yearly basis to enable collection of the necessary data to estimate the item parameters on the same scale as the other in the bank.

Examinations at the End of Secondary Education

At the end of secondary education, the students take a set of final examinations in a number of subjects that they selected earlier. After passing these examinations, they gain access to different forms of further education. The final examinations in most subjects are divided into two parts: a school examination and a national examination. The elements that are tested in each examination are specified in the examination syllabus, which is approved by the *College voor Examens* (CVE) (English translation: Board of Examinations, an arm's length body of the Ministry of Education). The CVE is also responsible for the level setting of the examinations. In the majority of examinations, the level-setting procedure is dominated by the information obtained using the assumption of random equivalent groups. Some other examinations have a small number of candidates; consequently, there is insufficient information about the performance of candidates. In such cases, a content judgment is used as the basis for the level setting. More elaborate data collection provides extra information for specific examinations considered central to the examination system. These include examinations in basic skills (Dutch language and mathematics), modern languages (English, French, and German), science (physics, chemistry, and biology), and economics. For these examinations, the additional data are collected using a pretest or posttest design (Alberts, 2001; Béguin, 2000). In these designs, parts of past and future examination forms are combined into tests that are administered as a preparation test for the examination. In other instances, the data are collected in different streams of education. Based on the collected data and using a Rasch model, the new examination is linked to an old form of the test. In this way, the standard on the new form can be equated to the standard on the old form.

The amount of data collected in the pretest or the posttest design is relatively limited due to restrictions on security of the items. Consequently, the equated score is provided with a confidence interval. As input to the level-setting meeting, the results of the above linking procedure are combined with the results of linking based on an assumption of random equivalent groups from year to year and, in some cases, content-based judgements about the difficulty level of the examinations.

End of Primary School Test

At the end of primary education, schools are obliged to collect objective information about the most appropriate type of secondary education for students. Most of the schools (about 85%)

apply a test for this purpose called the *Eindtoets Basisonderwijs* (Cito, 2012; Van der Lubbe, 2007), whereas the remainder apply other tests and assessments.

The Eindtoets Basisonderwijs contains a compulsory section composed of 200 multiple-choice items on the Dutch language, as well as on arithmetic and study skills, and a voluntary section composed of 90 items on history, geography, and science. Each year, a new form of the test is constructed that contains only new items, and the results are linked to those of the previous year's test. Three linking procedures based on different sources of information are used in the standard setting in the Eindtoets Basisonderwijs. The linking procedures are based on the following:

1) Pretest data in which the pretest forms combine items from multiple test forms of different years in a complex incomplete design
2) Anchor data that are collected, using an internal anchor embedded within the test forms of a sample of approximately 3,000 pupils taking the test, and noting that the anchor counts to the final score of these pupils
3) An assumption of random equivalent groups based on a sample of 1,800 schools that participated in the test for the past four years and in which no large shifts in the performance or in the size of the school occurred in this period

In the *pretest equating*, a multidimensional equating procedure is applied in which each of the 13 domains in the test is modelled using a separate dimension that is correlated to the other dimensions (Béguin, 2000; Glas, 1989).

Equating based on the pretest data is relatively unstable due to the small sample size of approximately 600 pupils per item and its susceptibility to model imperfections when, for example, order effects or time constraints are present.

The conditions under which the test is administered pose an additional threat to the validity of the linking, i.e., often the stakes are low for the student because the outcome of the test will have less importance to the student than the actual test. The administration condition could have an effect on the motivation of the pupils and, therefore, result in bias of the linking.

In the pretest, we tried to diminish this effect by collecting data in such a way that the motivation of the students was similar for all the items. However, this step does not guarantee that motivational effects will have no effect on the pretest equating. Equating using an *anchor test* is far more robust due to its larger sample size and the fact that the test is administered

under high-stakes conditions. In addition, the design is simpler. Thus, potential problems associated with time pressure and order effects are more easily detected and addressed.

A potential drawback with the anchor test design is that the anchor becomes known, and this will result in an increase in performance on the anchor that does not reflect an actual increase in proficiency. Finally, equating based on an assumption of *random equivalent groups* is stable and robust if the composition of the population does not change over time. A potential drawback of this approach is that if changes in the performance of the population do occur, they will be ignored.

Over the past few years, significant weight has been given to the assumption of random equivalent groups (Van Boxtel, Engelen, & De Wijs, 2012). This is because the standard setting based on the other sources of information (pretest data and anchor information) is, to some extent, inaccurate, whereas the trends in performance over time are historically stable. For reporting at the student level, it is unlikely that the standard setting based on the assumption of random equivalent groups compromises the standard because year-to-year effects are very small compared with the differences among students. However, to ensure that potential trends in performance are detected, all other sources of linking data are analyzed and used as a check on the standard setting. The results of these analyses are published in a report on the performance at the system level, which is available to the public a few months later. This report includes the results based on the different sources of linking information, together with the confidence intervals and corrected for background variables.

The type of detail that can be provided in such a report cannot be incorporated in the operational standard setting because of time constraints and the impossibility of including uncertainty in the reported cut-score.

Maintaining Performance Using Random Equivalent Groups Equating Instead of Maintaining Standards Using Nonequivalent Group Designs?

A number of equating procedures have been described earlier in the chapter, with some examples provided from tests in the Netherlands. In the level setting of both the central examinations in secondary education and in the end of primary school test, it is considered crucial that the standards are maintained over time.

In contrast to this, operationally the level setting depends at least partly on random equivalent groups equating, which theoretically just maintains performance. The reason for this seemingly invalid procedure is that in these tests the expected difference in performance level

between the years is expected to be smaller than the standard error of the equating procedures used to maintain the standard.

As a consequence, maintaining performance will be expected to reduce the instability of the level setting in a single year by trading the potential instability of the equating procedure for the potential bias due to the assumption of random equivalent groups.

According to the argument above, random equivalent groups equating could theoretically be used as the only source of information for level setting for these tests. However, there is a drawback for the maintaining-performance-only approach: Over multiple years, the bias would accumulate if a trend in performance in the population would occur. Another drawback with this approach is that maintaining performance could potentially undermine trust in the assessment system if the public considers that this procedure leads to a decrease in performance.

To be able to respond to claims about decreasing performance, it is crucial that a trend in performance can be evaluated at the system level and that standards can be maintained at that level. Therefore, next to maintaining performance based on random equivalent groups equating, equating information using additional data (like pretest and anchor test) also needs to be available, such as is the case in the examinations in secondary education and the test at the end of primary education. According to the additional data, it is possible to report on trends in performance in a detailed and nuanced way.

For example, it is possible to publish results, together with a confidence interval, or to report on different sources of equating information that contradict each other. Operationally, reporting in this level of detail is possible only at the system level, because uncertainty about standards cannot be included in a practical way in reports at pupil and school levels. In practice, some situations will present a difference between the cut-scores based on random equivalent groups equating and used for pupils and schools and the reported results of the performance in relation to the standards that include more sources of information.

In these cases, a correction will need to be made to the basis used for comparison in the test administered in the following year. This will prevent the accumulation of differences over the years from compromising the standard.

In summary, using a system based on maintaining performance (using random equivalent groups equating), combined with a number of equating procedures that are not necessarily all used as direct input in level setting, seems operationally to be the best option in circumstances where the expected differences in performance from year to year are smaller than the expected standard error of the equating procedures. The result from the equating procedures will be used in analyses at the system level to report on trends in performance in a detailed and nuanced way.

Although this procedure will potentially lead to a (probably small) deviation from the standard at the individual and school levels each year, the use of this approach over a number of years will not necessarily result in the accumulation of bias.

References

Alberts, R. V. J. (2001). Equating exams as a prerequisite for maintaining standards: Experience with Dutch centralised secondary examinations. *Assessment in Education: Principles, Policy & Practice, 8,* 353-367.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Béguin, A. A. (2000). *Robustness of equating high-stakes tests.* PhD thesis, University of Twente, Enschede.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic Press.

Cito (2012).Handleiding Eindtoets Basisonderwijs [Manual End of Primary School Test], Arnhem, the Netherlands: Cito.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15,* 20-31.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* (Doctoral Thesis.) Enschede: University of Twente.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (pp. 433–470). Westport, CT: American Council on Education.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item nonequivalent groups equating design. *Applied Psychological Measurement, 26*, 3-24.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131-143.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating* (2nd ed.). New York: Springer.

Levine, R. E. (1955). Equating the score scales of alternative forms administered to samples of different ability. *Research Bulletin 55-23*, Educational Testing Services, Princeton,NJ

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed. ,pp. 221-262). New York: American Council on Education and Macmillan.

Van Boxtel, H., Engelen, R., & De Wijs, A. (2012). *Verantwoording van de Eindtoets Basisonderwijs 2010.* Arnhem: Cito.

Van der Lubbe, M. (2007). *The End of Primary School Test (better known as Citotest)*. Paper presented at the 33rd annual conference of the International Association for Educational Assessment, September 16-21, Baku, Azerbaijan.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1994). *OPLM: Computer program and manual.* Arnhem: Cito.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press University of Chicago.